1 Total variation

Total-variation (TV) regularization reconstructs an image u from a noisy image f as the minimizer of the following functional:

$$F(u) = \int_\Omega |\nabla u| + \frac{\lambda}{2}\int_\Omega |u - f|^2. \tag{1}$$

Here, Ω is the image domain, generally a rectangle, the first term is $TV(u)$, the total variation of u, the second term is the data fidelity term, and λ is the regularization parameter, which controls the relative importance of the two terms.

For functions u that are not differentiable, we need a weak definition of $TV(u)$:

$$TV(u) = \max_g \int_\Omega u\nabla \cdot g, \tag{2}$$

where the maximization is over two-dimensional-valued functions $g \in C_c^1(\Omega)$ such that $\|g\|_\infty \leq 1$, where $\|g\|_\infty = \|(g_1, g_2)\|_\infty = \|\sqrt{g_1^2 + g_2^2}\|_\infty$. That this equals $\int |\nabla u|$ when u is smooth is a consequence of integration by parts, plus the elementary fact that $|x| = \max_{|v|\leq 1} xv$ as well as $\max_{|v|\leq 1}(-xv)$. The result of this is that $|\nabla u|$ is in general a measure, though one that can be identified with a function in many cases.

The functions $u \in L^1(\Omega)$ satisfying $TV(u) < \infty$ are said to have bounded variation, and form the space $BV(\Omega)$. Adding the L^1 norm to TV gives a norm which makes BV a Banach space. This is the domain of F (since $BV \subset L^2$). This functional can be shown to be lower semicontinous and is strictly convex, which guarantees that it has a minimizer and that the minimizer is unique.

1.1 Euler-Lagrange equation

To minimize F, we set an appropriate sense of derivative equal to zero. This sense is the first variation, also known as the Gateaux derivative, which we will write as simply F'. First, we fix arbitrary $v \in BV$ (a "test function") and compute the directional derivative:

$$D_vF(u) = \int \frac{\nabla u}{|\nabla u|} \cdot \nabla v + \lambda \int (u - f)v. \tag{3}$$

This clearly requires $\nabla u \neq 0$ almost everywhere to make sense. If you go through the computation in careful detail, you find that the limit doesn't exist for every $v \in BV$; for example, $\int |\nabla v|^2$ needs to be finite. This makes the derivatives involved less than completely rigorous, something usually shrugged off due to the eventual implementations being on finite-dimensional subspaces.

We then obtain $F'(u)$ as the "function" satisfying $D_v F(u) = \langle F'(u), v\rangle$ for any test function, though it is even less likely to truly be a function than $|\nabla u|$. We thus use integration by parts to isolate v:

$$D_v F(u) = \int \left(-\nabla \cdot \frac{\nabla u}{|\nabla u|} + \lambda(u - f) \right) v, \tag{4}$$

giving

$$F'(u) = -\nabla \cdot \frac{\nabla u}{|\nabla u|} + \lambda(u - f). \tag{5}$$

The Euler-Lagrange equation for the problem of minimizing F is obtained by setting the right side of the previous equation to zero. The boundary term from the integration by parts can be dropped if we assume that the normal derivative of u at the boundary is zero. This leads to Neumann boundary conditions in PDE minimization methods.

1.2 Linear operator inversion

One use of total-variation minimization is to regularize the inversion of linear operators. This is especially useful in the case that the operator has a smoothing effect, in which case its inverse will amplify noise. Examples include the following:

operator	inverse	purpose
identity	identity	denoising
convolution	deconvolution	deblurring
Abel transform	Abel inverse	radiographic inversion of cylindrically-symmetric objects
antidifferentiation	differentiation	numerical differentiation of noisy data

Suppose the linear operator is T. In each of the above cases, T is invertible and bounded (on the appropriate space), but in all but the first case the inverse operator is unbounded.

We incorporate T into the functional:

$$F(u) = \int_\Omega |\nabla u| + \frac{\lambda}{2}\int_\Omega |Tu - f|^2. \tag{6}$$

We thus are measuring data fidelity in the “data space” but regularizing in the “object space.” This terminology arises in the case of radiography, where f is a radiographic image (the data), and u describes an object to be reconstructed from the radiograph.

The resulting Euler-Lagrange equation is

$$0 = -\nabla \cdot \frac{\nabla u}{|\nabla u|} + \lambda T^*(Tu - f). \tag{7}$$

Backing up one step, this is because in the directional derivative, the effect of the chain rule is to pass the operator on to v:

$$D_v\left((Tu - f)^2\right) = 2(Tu - f)Tv; \tag{8}$$

we then isolate v by moving T over to T^* on the other side of the L^2 inner product. All this clearly depends on T being linear.

To avoid division by zero, the $|\nabla u|$ in the denominator is usually replaced by $\sqrt{|\nabla u|^2 + \epsilon}$ for some small $\epsilon > 0$. This not only adds an extra parameter to worry about (though one that is not terribly sensitive), but means that the computed minimizer will never be the true minimizer of F.

1.3 Gradient descent

Now we turn to the question of how to find the minimizer of F. The simplest approach is gradient descent. We can formulate this as the PDE $u_t = -F'(u)$. This is solved iteratively by starting with some u_0, then solving

$$\frac{u_{n+1} - u_n}{\Delta t} = -\nabla \cdot \frac{\nabla u_n}{|\nabla u_n|} + \lambda T^*(Tu_n - f) \tag{9}$$

for some small Δt. A more sophisticated approach is to find the best Δt for each n (“at each timestep”), as that which gives the greatest decrease in F.

Gradient descent is simple to understand and implement, but converges slowly. The closer u_n is to the minimum, the slower the convergence.

1.4 Lagged diffusivity

An algorithm that converges faster is lagged diffusivity. It is obtained from the Euler-Lagrange equation by "lagging" the nonlinear term:

$$0 = -\nabla \cdot \frac{\nabla u_{n+1}}{|\nabla u_n|} + \lambda T^*(Tu_{n+1} - f). \tag{10}$$

Given u_n, the equation is linear in u_{n+1}. This iteration has been proven to converge (assuming T has trivial nullspace). If we call G the operator $-\nabla \cdot \frac{\nabla}{|\nabla u_n|} + \lambda T^* T$, then u_{n+1} is the solution to the equation $G(u_{n+1}) = T^* f$. Supposedly, one gets better numerical performance from instead solving $u_{n+1} = u_n - G^{-1}(G(u_n) - T^* f)$.

1.5 Other data fidelity terms

The L^2 data fidelity term is most appropriate when the noise in f is additive, Gaussian noise. In the case of radiography or when f is obtained by counting particles, f will obey a Poisson distribution. The noise is neither additive nor multiplicative, and is referred to a "Poisson noise." In this case, the best data fidelity term is $\int(u - f \log u)$. It is unusual in that it doesn't come from a norm, and can take on negative values. It is, however, bounded below, with the minimum being when $u = f$. With an operator, it would be $\int(Tu - f \log(Tu))$. The Euler-Lagrange equation becomes

$$0 = -\nabla \cdot \frac{\nabla u}{|\nabla u|} + \lambda T^* \left(1 - \frac{f}{Tu}\right). \tag{11}$$

1.6 Dual algorithms

We are minimizing a functional defined in terms of $TV(u)$, which itself is defined in terms of a maximum. Convexity arguments allow one to interchange the order of the optimizations. The inner minimization has an Euler-Lagrange equation that can be solved explicitly and subsituted back in. The outer maximization then becomes the optimization problem to solve. The condition that $\|g\|_\infty \leq 1$ becomes a constraint, but one for which the Lagrange multiplier can be solved for. One then arrives at a new algorithm that solves for g, from which one can obtain u. Its chief advantage is there is no concern with division by zero, and no need to introduce ϵ. It thus also gives a minimizer to F itself instead of an approximation.

1.7 Radiographic corrections

In the case of Abel inversion, the operator T maps object density (given by u) to "areal density," mass per unit area. This is closely related to what one would expect to measure with a radiograph, namely transmission (such as, how much X-ray intensity goes through the object to the detector). One can try to translate radiograph transmission into an areal density, which becomes f for the total-variation regularized Abel inversion process. However, there are reasons to try to use the radiograph intensity itself for f, or at least something closer than areal density.

Another way to look at this is that the Abel transform is a linear approximation to the radiographic process, but we would like to use a better, nonlinear approximation. An improvement is to compose with T a function of the form

$$h(t) = e^{\gamma_1 t}\big(1 - e^{\gamma_2/t}\big), \tag{12}$$

where γ_1 and γ_2 are constants. On the one hand, $h \circ T$ is nonlinear. On the other hand, it is differentiable, and the Euler-Lagrange equation can still be written explicitly.

One framework for dealing with algorithms involving nonlinear functions and/or novel data fidelity terms is to work in terms of a general data fidelity term, $\int D(u, f)$. Sometimes this makes things clearer.

1.8 Staircase reduction

One problem with TV-based methods is staircasing. This has to do with the fact that $TV(u)$ depends only on how much u varies, and not on how. In particular, discontinuities are not penalized. For many purposes, this is a good thing, as it allows sharp edges to be obtained. However, when f is noisy, in regions where u should vary smoothly, the obtained u will often instead resemble a step function.

A fix for this is to decompose u into $u_1 + u_2$, and minimize instead

$$\alpha_1 \int |\nabla u_1| + \alpha_2 \int |\Delta u_2| + \int D(u_1 + u_2, f). \tag{13}$$

Note that λ is replaced by the two parameters α_1, α_2. One then gets two Euler-Lagrange equations, one for u_1 and one for u_2, the latter involving 4th-order derivatives. One can then use gradient descent, alternating between the two equations with each (or every several) timestep(s). A dual version of the algoritm also exists. This has been done with L^2 data fidelity term as well as one other term I won't get into, but not Poisson or general terms. A simpler version uses $|\Delta u_2|^2$ instead of $|\Delta u_2|$.

2 Possible Tasks

There are several issues to explore, both computational and theoretical. We could work on any one or combination of the following:

1. implement gradient descent for Abel inversion with nonlinear corrections, with an L^2 data fidelity term, Poisson data fidelity term, or both. This would at least be good enough for a technical report that would find many interested LANL readers, and may be publishable.

2. implement staircase reduction for Abel inversion, with L^2 or Poisson data fidelity terms, with or without nonlinear correction.

3. formulate analogs of lagged diffusivity for other data fidelity terms, and try to prove that they converge. Success would mean a paper. Success for a general data fidelity term, even allowing only linear operators, would be a strong paper.

4. numerical implementation of the previous.

5. other?

2.1 Notes and references

I'm including several papers. It certainly isn't necessary to read every one or pay attention to every detail.

- two papers on TV-Abel inversion, asaki-2005-abel.pdf and asaki-2005-abel2.pdf. The first pays more attention to the theory, the second to the implementation.

- a paper on TV-regularized differentiation, chartrand-2005-numerical.pdf. A nice example of regularizing a noise-amplifying process, and includes a lagged diffusivity implementation (though briefly).

- a paper with the Poisson data fidelity term, le-2005-variational.pdf. It includes a derivation of how the data fidelity term arises from the assumption of Poisson noise.

- a lengthy paper dealing with staircase reduction, and other topics, cam05-28.pdf.

- a very preliminary draft of a paper concerning dual algorithms and general data fidelity, general.pdf. It is incomplete and not well written yet. It is in the context of the dual algorithm approach above, but it illustrates how one can go about dealing with a general data fidelity term. Also, if you ignore how G comes about, it gives an approach for proving convergence in a general context, which may be applicable to the lagged diffusivity algorithm.

- a paper giving the dual algorithm referred to at length in the previous, chambolle.pdf. I include it in case it helps make the previous make more sense.

- there are no good references for a proof of the convergence of the lagged diffusivity algorithm, even for L^2 data fidelity with no operator. Proofs exist, but are very ad hoc and hard to follow. All are very particular to the L^2 case. I imagine a successful proof coming more along the lines of the approach in general.pdf.

- Curt Vogel's book gets into the details of numericaly implementing the lagged diffusivity algorithm. I'll leave my copy on my desk in case you don't have one.